STAT0006 ICA 3

Group 19

Student numbers: 23168744, 22069847, 22031241, 23011899

Part 1: Normal linear model

Analysis of property prices

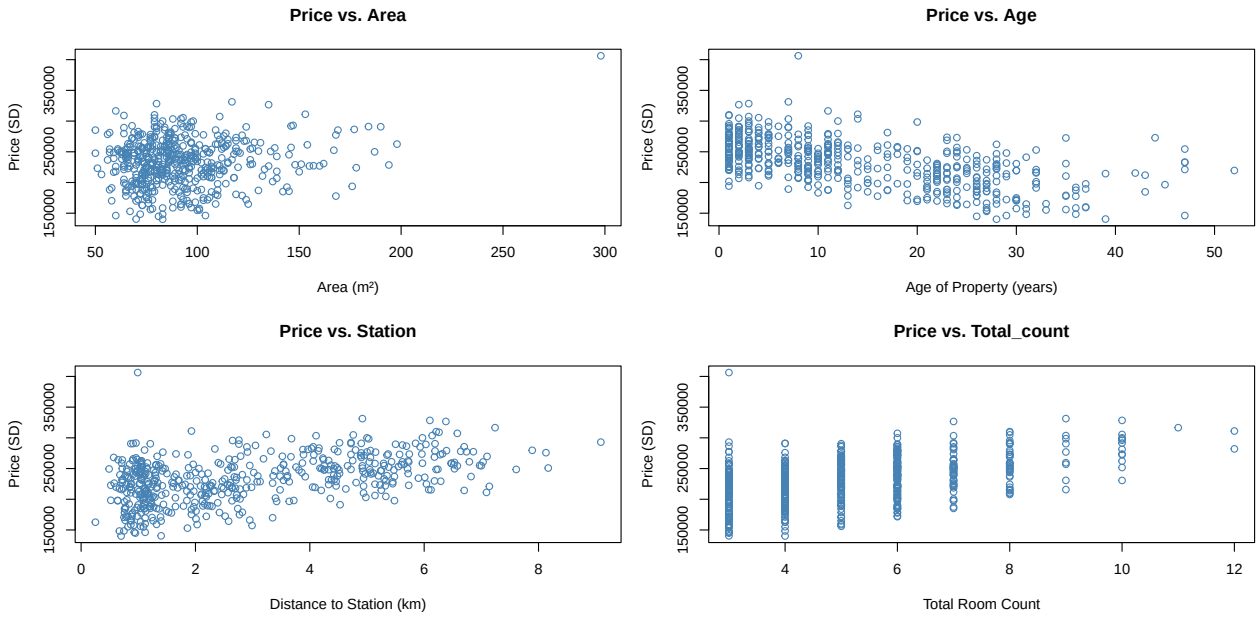
Introduction to the data

A market research on residential property prices in the city of Statsville is carried out by a property development company. A random sample dataset of the properties sold in the last two years has been collected. The company wants to understand the underlying factors which influence the prices that properties sell for. The data is available in prices.csv.There are no missing entries in this dataset. The average price of properties is 232,590 SD with an interquartile range of 49257.50 SD. The prices range of from a minimum of 140,115 SD to a maximum of 406,305 SD.

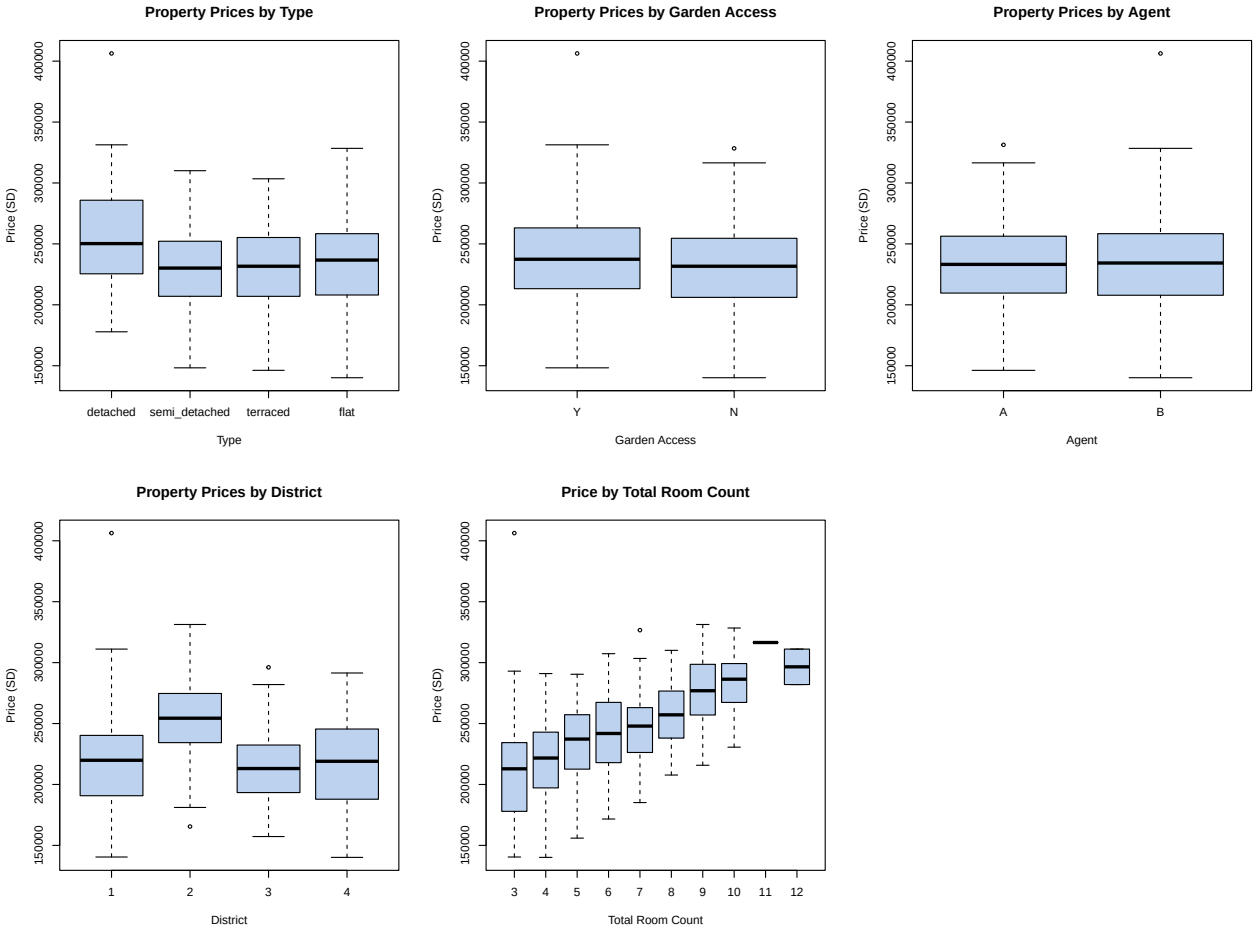
There are 500 observations with 11 variables provided:

- `price` (in SD)
- `bedroom_count` (number of bedrooms in the property)
- `other_count` (the number of other rooms, excluding bedrooms)
- `total_count` (the total number of rooms in the property)
- `age` (age of property in years)
- `district` (categorised as 1, 2, 3, 4 by the company)
- `type` (either detatched, semi_detached, terraced or flat)
- `garden` (Y or N indicating whether the property has access to a garden)
- `area` (in square metres)
- `station` (distance to the nesarest train station in km)
- `agent` (either estate agent A or who handled the sale)

Scatter plots reveal the linear relationship between price and four variables: `age`, `area`,`total_count`, and `station`. All variables seem highly related with price except`area`, with `age` and `total_count`show a positive trend, indicating that older properties and more room number tend to have higher prices. Conversely, station shows a negative trend, where properties farther from train stations generally have lower prices. In contrast, `area` does not display a clear linear relationship with price, suggesting that a transformation may be needed or it may have little to no significant impact. Since `total_count` combines information from `bed_count` and `other_count`, it provides a comprehensive view of how room numbers influence price, so we are only looking into the effect of total_count on price in this part.



We compare the remaining variables' effects on tips by making the boxplots (the bolded black lines represent the median amount of tips under specific variable). Detached property type and properties in District 2 seem to have a significant impact on the increase in price. Agent choice and garden access appear to have minimal influence on price, as the medians of the categories for both plots are nearly identical. The `total_count` shows a strong positive relationship with price, as properties with more rooms are consistently valued higher, making it a key factor in determining property prices. Note that a linear relationship between total_count and price is clearly captured which suggest that treating the covariate as numeric may be more appropriate.



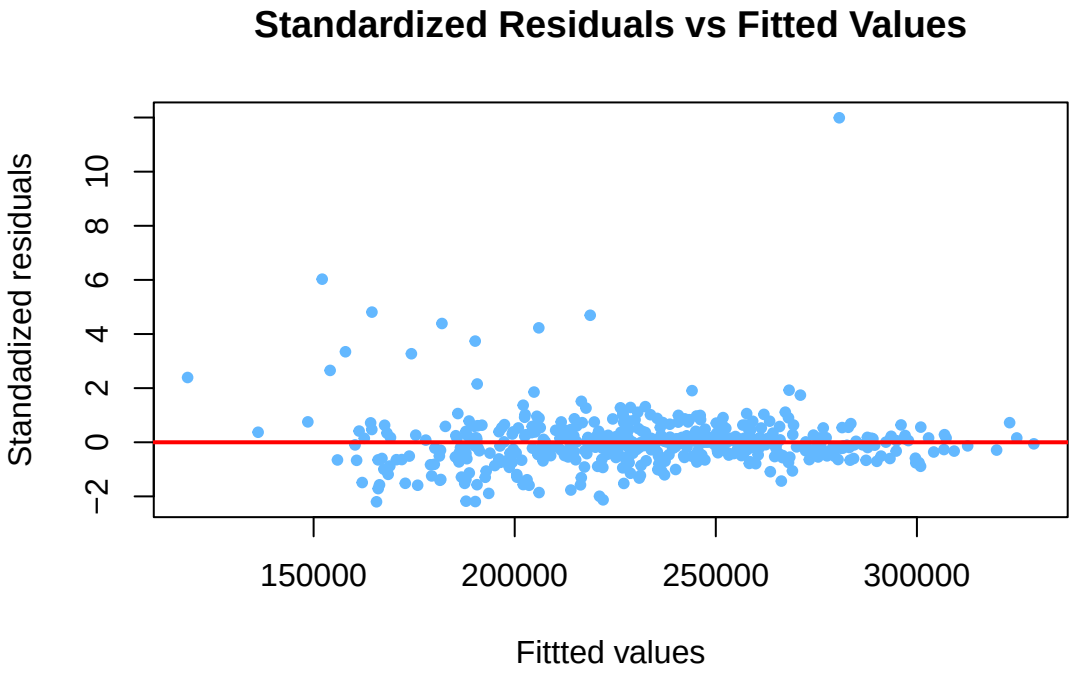
Some prices appear unusually high compared to the majority in the boxplots. Additionally, there is an imbalance in sample sizes across certain variable categories,especially in `type` and `distract` where the maximum sample size can be up to four times larger than the minimum. This imbalance could lead to reduced accuracy or introduce bias in the model's predictions for smaller sample categories.

In conclusion, `age` `total_count`and `station` have significant relationships with price but with opposite trend. `district`, and `total_count` are strong predictors of price: detached properties, Districts 1, and properties with more rooms commanding higher prices. Agent and garden access shows no significant effect on price. However, sample imbalances in type and district and outliers in high-price distributions might potentially affect accuracy of result need further investigation and verified.

Model building

Handling influential observations

To start, it's crucial to examine a data point at a price that is 49257.50 SD. This data point could potentially be an outlier or a leverage point resulting from input errors, an extreme isolated event, or a model inadequacy. Upon further investigation of the data point, the number of bedrooms and other rooms appears minimal in comparison to properties in the higher price range. Although the property of interest has significant difference in the area and distance from the nearest train station relative to other expensive properties, there are also properties with more affordable sale prices that share similarities in area and distance.



To futher justify this. we then look at a plot of standardized residuals against fitted values. The extremely high standardized residuals indicate that this data point is an outlier. Since there are no other signs suggesting it is an input error, the data seems to be legitimate. As a result, we will not eliminate this outlier.

Linearity check

Initially, we construct a regression model that includes all the covariates, which we will refer to as Model 1. The summary of Model 1 is presented as follows.

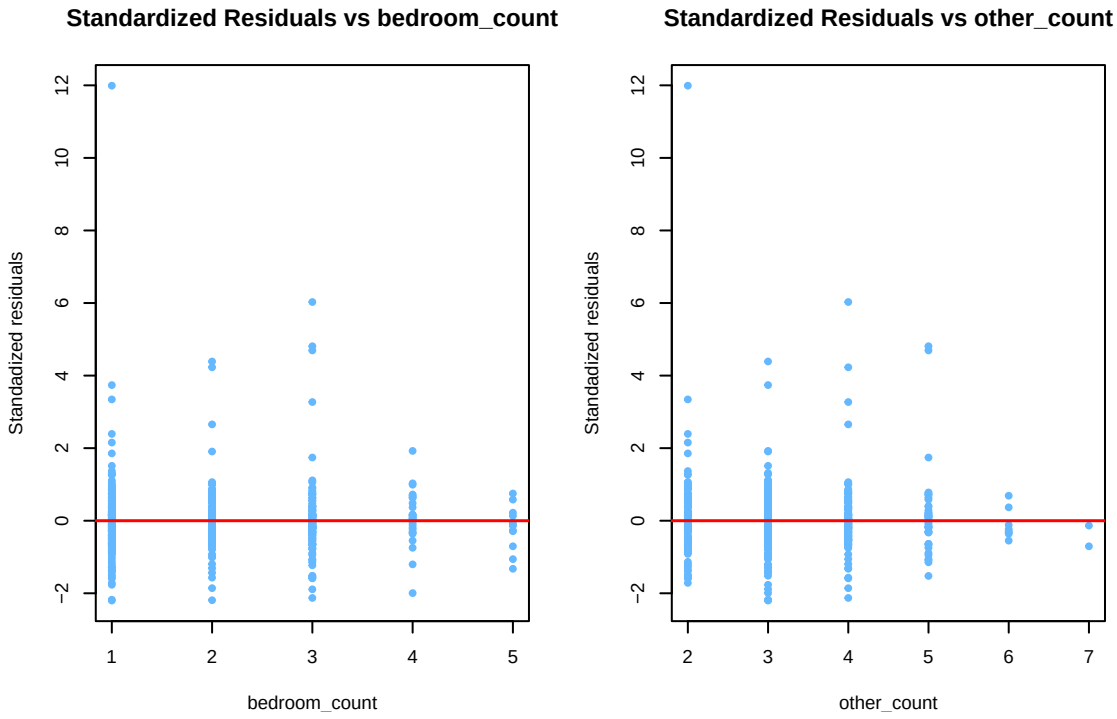
```
##
## Call:
## lm(formula = price ~ ., data = prices)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25490  -5236   -814   3043 125602
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       159652.89    6171.89  25.868  < 2e-16 ***
## bedroom_count      14592.83     634.00  23.017  < 2e-16 ***
## other_count         6171.44     663.93   9.295  < 2e-16 ***
## total_count              NA         NA      NA       NA
## age                -1872.76      46.36 -40.396  < 2e-16 ***
## district2          20769.50    2521.21   8.238 1.64e-15 ***
## district3         -13643.11    2039.93  -6.688 6.23e-11 ***
## district4           1717.50    1733.29   0.991    0.322
## typesemi_detached   3519.61    3093.67   1.138    0.256
## typeterraced         406.77    2902.95   0.140    0.889
## typeflat            1197.63    3603.07   0.332    0.740
## gardenN            -9669.03    1404.02  -6.887 1.77e-11 ***
## area                 351.27      32.99  10.649  < 2e-16 ***
## station             5165.98     570.60   9.054  < 2e-16 ***
## agentB              -695.35    1058.43  -0.657    0.512
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11710 on 486 degrees of freedom
## Multiple R-squared:  0.9066, Adjusted R-squared:  0.9041
## F-statistic: 362.9 on 13 and 486 DF,  p-value: < 2.2e-16
```

NA values for the coefficients of total_count in the model summary indicates perfect multicollinearity. In this case, total_count is a sum of bedroom_count and other_count. Comparison between keeping bedroom_count and other_count or total count needs to take place to reduce redundancy.

Additionally, transformations are implemented to covariate area but non improves the correlation of area and price. However, the p-value of area shown in the summary is significantly low addressing the earlier concern about the absence of a linear relationship. This because the regression model adjusts for these other variables, meaning the p-value reflects the unique contribution of area to explaining price after accounting for other predictors.

Before proceeding with model building, it's essential to assess whether the model follows a linearity assumption by examining standardized residuals in relation to each numerical covariate. Below are the plots of standardized residuals against bedroom_count and other_count.



The residuals seem to be evenly distributed for both bedroom_count and other_count, with a slight clustering on the right side but no distinct systematic trend. This indicates that it is appropriate to treat bedroom_count and other_count as numeric variables. At this stage, we have also reviewed plots for each numeric covariate, allowing us to proceed to the next step.

Models tesing

Two nested models, subset of Model 1, omitting the covariates with large p-values showed in the summary. Model 2.1, with only low p-value variables: bedroom_count, other_count, age, district, garden, area, station, and Model 2.2 where bedroom_count and other_count are replaced by total_count. The summary is shown below to investigate wheter bedroom_count and other_count or total_count would give the better view of the data.

```
##
## Call:
## lm(formula = price ~ bedroom_count + other_count + age + district +
##     garden + area + station, data = prices)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25862  -5221   -876   3528 126360
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   161710.60    3540.65  45.673  < 2e-16 ***
## bedroom_count  14533.96     632.05  22.995  < 2e-16 ***
## other_count     6218.26     660.87   9.409  < 2e-16 ***
## age            -1872.84      46.14 -40.587  < 2e-16 ***
## district2      20586.43    2502.46   8.226 1.75e-15 ***
## district3     -13597.25    2035.56  -6.680 6.51e-11 ***
## district4       1757.60    1724.26   1.019    0.309
## gardenN        -9989.89    1216.66  -8.211 1.96e-15 ***
## area             339.33      22.19  15.294  < 2e-16 ***
## station         5179.20     569.58   9.093  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11700 on 490 degrees of freedom
## Multiple R-squared:  0.9059, Adjusted R-squared:  0.9041
## F-statistic:   524 on 9 and 490 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = price ~ total_count + age + district + garden +
##     area + station, data = prices)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29250  -6510  -1042   4792 129122
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 157591.11    3669.44  42.947  < 2e-16 ***
## total_count  10491.69     302.06  34.734  < 2e-16 ***
## age          -1859.18      48.42 -38.396  < 2e-16 ***
## district2    21126.23    2627.02   8.042 6.67e-15 ***
## district3   -13767.13    2137.70  -6.440 2.85e-10 ***
## district4     1435.52    1810.29   0.793    0.428
## gardenN     -10350.67    1276.70  -8.107 4.16e-15 ***
## area           328.90      23.25  14.145  < 2e-16 ***
## station       5029.99     597.80   8.414 4.35e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12290 on 491 degrees of freedom
## Multiple R-squared:  0.896,  Adjusted R-squared:  0.8943
## F-statistic: 528.6 on 8 and 491 DF,  p-value: < 2.2e-16
```

The R-squared of Model 2.2 is 0.896 meaning 89.6% of the proportion of variation of the sale price is explained by the predictors which is just slightly smaller Model 2.1 of R-squared being 0.9059. Given that both models exhibit relatively similar R-squared values, we need to choose one of them to avoid redundancy. To achieve this, we will conduct an F-test to evaluate

An F-test is carried to compare Model 1 and Model 2.1 to test the hypothesis that we can omit total_count, type and agent entirely.

```
## Analysis of Variance Table
##
## Model 1: price ~ bedroom_count + other_count + age + district + garden +
##     area + station
## Model 2: price ~ bedroom_count + other_count + total_count + age + district +
##     type + garden + area + station + agent
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    490 6.7121e+10
## 2    486 6.6595e+10  4 526684994 0.9609 0.4286
```

Under the null hypothesis, we prefer the model without total_count, type and agent, Model 2.1. The results of the F-test show a test statistic of 0.9609 and a p-value of 0.4286. Since this p-value exceeds 0.005, we lack strong evidence to reject the null hypothesis. This indicates that there is no compelling reason to believe that the nested model (Model 2.1) is a less effective representation of the data than the full model. Therefore, we conclude that the simpler model remains a valid option for adequately describing the data.

Now, we test Model 2.2 against the initial full model.

```
anova(model_22,model_1, test="F")
```

```
## Analysis of Variance Table
##
## Model 1: price ~ total_count + age + district + garden + area + station
## Model 2: price ~ bedroom_count + other_count + total_count + age + district +
##     type + garden + area + station + agent
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    491 7.4188e+10
## 2    486 6.6595e+10  5 7592755201 11.082 3.999e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Under the null hypothesis, we prefer the model without bedroom_count, other_count, type and agent, Model 2.2. With the test statistic for F-test being 11.082 and p-value of 3.999e-10 which is very small. We conclude that there is an evidence to reject null hypothesis suggesting that the nested model is not as good a description of the data as the full model.

After comparing the results from both tests, it is evident that we should proceed with Model 2.1. This also provides the stakeholder with more insights into how each bedroom_count and other_count influences the price.

At this stage, it is essential to examine collinearity after eliminating the redundant variable. We need to ensure that we investigate any other variables that may be highly correlated if they emerge in the analysis.

```
##                    GVIF Df GVIF^(1/(2*Df))
## bedroom_count 1.597967  1        1.264107
## other_count   1.584453  1        1.258751
## age           1.010572  1        1.005272
## district      4.691895  3        1.293872
## garden        1.250837  1        1.118408
## area          1.285537  1        1.133815
## station       4.596022  1        2.143834
```

The threshold of Varaince Inflation Facotr(VIF) to be considered that the model has problems estimating the coefficient is being larger than 5. The results show all the VIF being between 1 and 5 suggesting moderate collinearity which is generally acceptable.

Revisiting the summary of Model 2.1, we observe that the p-values for the dummy variables representing the districts having district 1 as the reference—indicate that districts 2 and 3 have low p-values, whereas the p-value for district 4 is comparatively higher. We will develop a nested model based on Model 2 called Model 3 that excludes the district covariate. An F-test will be conducted once more.

```
## Analysis of Variance Table
##
## Model 1: price ~ bedroom_count + other_count + age + garden + area + station
## Model 2: price ~ bedroom_count + other_count + age + district + garden +
##     area + station
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    493 1.0584e+11
## 2    490 6.7121e+10  3 3.8721e+10 94.224 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Under the null hypothesis, we prefer the model that exclude district entirely. With the test statistic for F-test being 94.224 and p-value of less than 2.2e-16 which is very small suggesting that there is a significant evidence to reject null hypothesis. As a result, we do not omit district from the model.

In the end, we experimented with various combinations of interactions, and the interaction that enhances the model is between bedroom_count and area. This is logical, as a higher number of bedrooms is likely to correspond with an increase in the total area.

```
## Analysis of Variance Table
##
## Model 1: price ~ bedroom_count * area + other_count + age + district +
##     garden + station
## Model 2: price ~ bedroom_count + other_count + age + district + garden +
##     area + station
##   Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
## 1    489 6.6347e+10
## 2    490 6.7121e+10 -1 -774801562 5.7106 0.01724 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

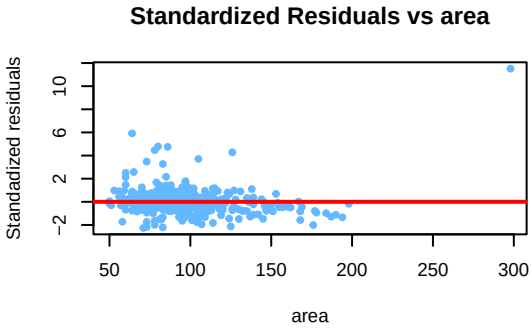
This is the comparison of Model 4 with Model 2.1. According to the null hypothesis, Model 4, which includes the interaction between bedroom_count and area, is preferred. The F-test yielded a statistic of 5.7106 with a p-value of 0.0172, which is below the 0.05 threshold. This indicates that we do not have sufficient evidence to reject the null hypothesis. Therefore, we incorporate the interaction into our model.

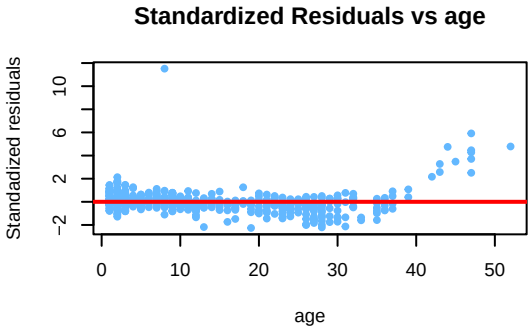
Model checking for final chosen model

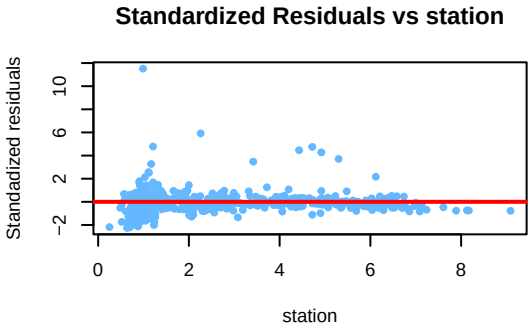
The final model is Model 4 with the covariates being an interaction between bedroom_count and area, other_count, age, district, garden and station. We will now perform a final check for the model including departures from linearity, homoscedasticity, normality and independence.

Next, we need to verify whether the assumptions regarding the error terms are met. The primary assumption necessary for estimating the regression coefficients is linearity. The other assumptions are important for facilitating hypothesis testing and constructing confidence intervals.

Departure from Linearity
We see if the expected value of the response is a linear combination of covariates.

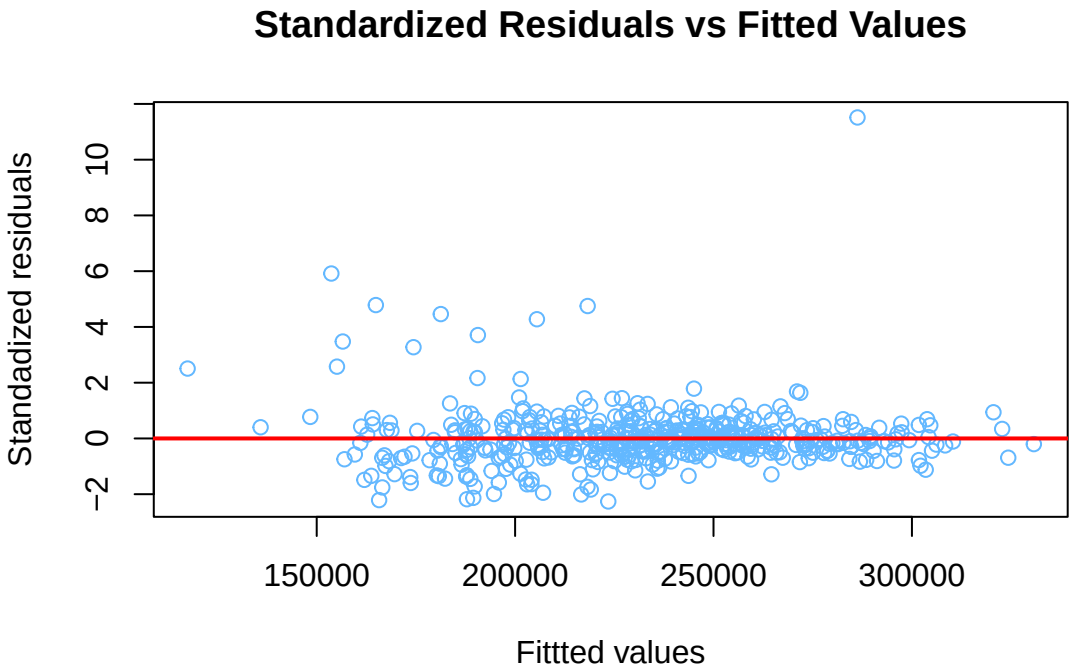






Linearity check has been shown for bedroom_count and other_count. The scatter plots for the rest of the numeric covariates: area, age and station also show relatively even spread as well. With all being said, the linear assumption is hold for this final model.

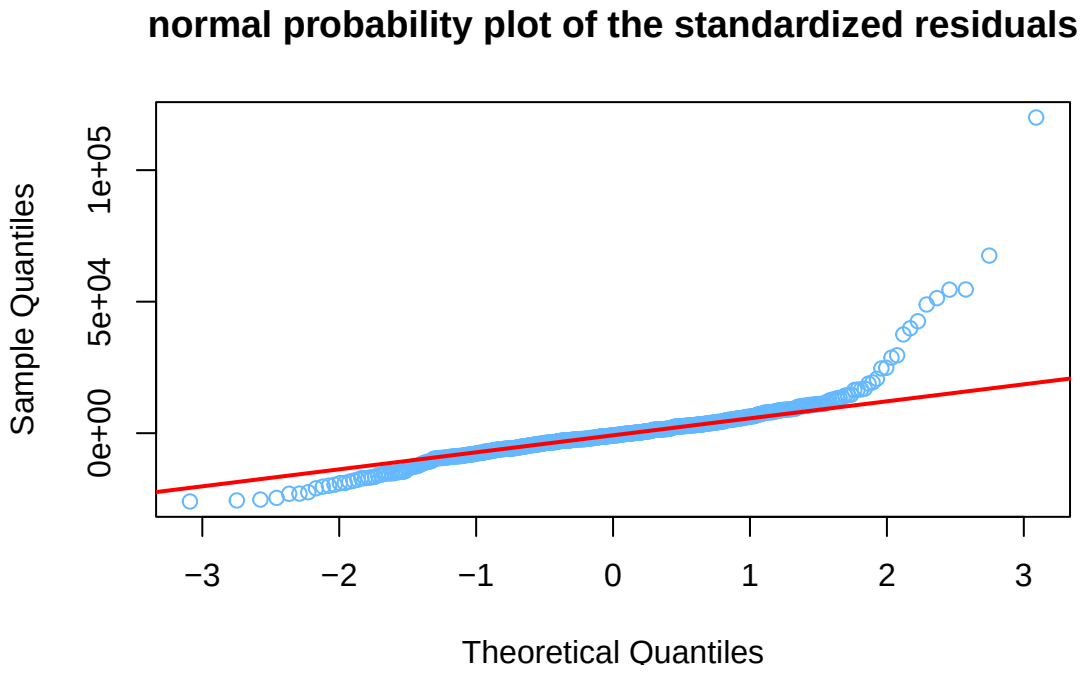
Departure from homoscedasticity



The plot is fairly evenly spread, although there is a little bit of clustering noticed closer arounf the zero line. Overall, their is no significant evidence against homoscedasticity assumption.

Departure from normality

We plot the standardized residuals against the fitted values to examine the assumption of homoscedasticity. The is to determine whether the residuals are consistently spread across the fitted values. The resulting plot is shown below:



The plot indicates that the standardized residuals have a distribution with heavier tails than those of the standard normal distribution, providing evidence against the normality assumption. However, it does not make the model inadequate as the assumption is on errors.

Departure from independence

We implement Durbin-Watson test here to observe any presence of serial-correlation.

```
##
##  Durbin-Watson test
##
## data:  model_4
## DW = 1.9906, p-value = 0.4598
## alternative hypothesis: true autocorrelation is greater than 0
```

Durbin-Watson statistic is about 2 and p-value is greater than 0.05 suggesting no significant serial correlation. Therefore, the independence assumption is satisfied.

Conclusion

The final model is a linear regression that explains approximately 90.51% of the variability in property prices, including both continuous (bedroom_count, other_count, area, station, age), categorical (district, garden) variables. Diagnostic checks confirm that the model satisfies key assumptions, although residuals show mild departures from normality. The inclusion of interaction terms, such as between bedroom_count and area, improved model performance. Below is the summary of the final model.

```
##
## Call:
## lm(formula = price ~ bedroom_count * area + other_count + age +
##     district + garden + station, data = prices)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25960  -5165   -858   3553 120017
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        154570.55    4619.97  33.457  < 2e-16 ***
## bedroom_count       18914.93    1938.20   9.759  < 2e-16 ***
## area                  417.86      39.59  10.554  < 2e-16 ***
## other_count          6273.88     658.13   9.533  < 2e-16 ***
## age                 -1875.39      45.94 -40.826  < 2e-16 ***
## district2           20532.50    2490.62   8.244 1.55e-15 ***
## district3          -13570.67    2025.88  -6.699 5.80e-11 ***
## district4            1740.38    1716.04   1.014   0.3110
## gardenN            -10076.22    1211.39  -8.318 8.97e-16 ***
## station              5111.44     567.57   9.006  < 2e-16 ***
## bedroom_count:area    -48.07      20.11  -2.390   0.0172 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11650 on 489 degrees of freedom
## Multiple R-squared:  0.907,  Adjusted R-squared:  0.9051
## F-statistic: 476.7 on 10 and 489 DF,  p-value: < 2.2e-16
```

The covariates that appear to contribute in th estimated property price the most are bedroom_count, district and garden. The expected price of a property in Statville increases by 18914.93 SD with each one number of bedrooms increased holding all other covariates constant. Meanwhile, the expected price of a property in Statville increases by 20,532.50 SD in district 2 compared to district 1, but decreased by 13,570.67 SD in district 3 compared to district 1 holding all other covariates constant. Additionally, the expected price of a property without an access to a garden dropped by 10076.22 SD in comparison with a property with a garden access holding all covariates constant.

In general, the model offers important insights into the elements affecting property prices, highlighting the covariates that are statistically significant. These findings can guide decision-making in the property market.

Discussion of limitations

While the model satisfies most linear regression assumptions, mild departures from the normality of residuals were observed, as indicated by heavier tails in the Q-Q plot. Although this issue does not significantly affect the predictive capabilities of the model, it may impact the accuracy of statistical inference, such as p-values and confidence intervals.

Additionally, slight clustering patterns were observed in the residual plots, indicating potential structural issues within the data that the model does not fully capture. Together with the Q-Q plot departure from normality. An attempt was made to address these issues by applying transformations to the response variable and specific predictors. However, these transformations did not lead to significant improvements in the residual distribution or model fit. This suggests that the relationships between variables may be more complex than can be addressed by basic transformations. Implementing more advanced techniques could provide a better framework for capturing these patterns and improving the model's performance.

Lastly, The model does not account for several potentially important predictors that could influence property prices, such as neighborhood amenities, crime rates, or proximity to schools and parks. In the future analysis, incorporating such factors would improve the model's explanatory power and provide a more comprehensive understanding of property valuation.

Part 2: Generalised linear model OR Generalised additive model

Analysis of sales times

Data analysis

The manager of the property company aims to maximise the likelihood of a quick sale for a personal house priced at 200,000 SD. The updated dataset `times`, includes information on 500 past sales, indicating whether each property was sold within two months (Success: 1 or Failure: 0), along with its price in Statsland Dollars (SD) and the responsible agent (agent A or agent B). There are no missing values in this dataset, and there are no obviously unreasonable values either.

Before developing a predictive model, we examine the relationship between property price and the probability of a quick sale to determine if a parametric relationship exists. Figure 1 illustrates the proportion of quick sales across different price groupings, with each blue point representing a specific group. The analysis reveals that there is no clear linear relationship between price and the probability of a quick sale for either agent.

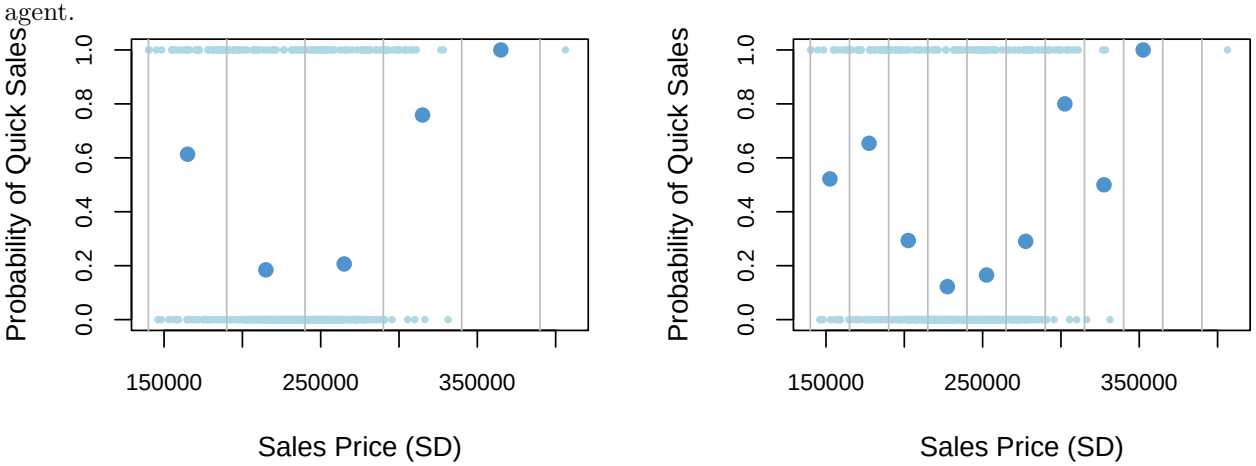


Figure 1: Plots showing the relationship between sale price and probability of quick sale in 50,000SD bins (left) and 25,000SD bins (right).

Since no clear parametric relationship exists between price and the probability of a quick sale, it is reasonable to take a non-parametric approach. The response variable, representing the success or failure of a quick sale, is binary and follows a binomial distribution within the exponential family. Therefore, a Generalized Additive Model (GAM) with a logit link function is appropriate to analyse the `times` dataset. This approach allows for the flexibility needed to capture potential non-linear relationships between price and the probability of a quick sale, while also incorporating the effect of the responsible agent. The R model summary for this model (times.gam) is given below:

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## quick_sale ~ s(price) + as.factor(agent)
##
## Parametric coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.2927     0.3274 -10.057   <2e-16 ***
## as.factor(agent)B    3.3818     0.3795   8.911   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Approximate significance of smooth terms:
##            edf Ref.df Chi.sq p-value
## s(price) 7.974  8.496  78.72  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.422   Deviance explained = 40.7%
## UBRE = -0.24348  Scale est. = 1         n = 500
```

From the model summary, we observe that the deviance explained is 40.7% and the R-squared is 0.422, which are not big numbers. While this indicates that the model captures some variability in the data, it does not necessarily imply a strong or satisfactory fit to the dataset.

Figure 2 provides some evidence against the assumption of normality, as the points in the QQ-plot show noticeable deviations from the reference line. Additionally, there is some indication of heteroscedasticity, as the variability does not appear to be constant across the range of fitted values. Thus, the model does not look like doing a great job.

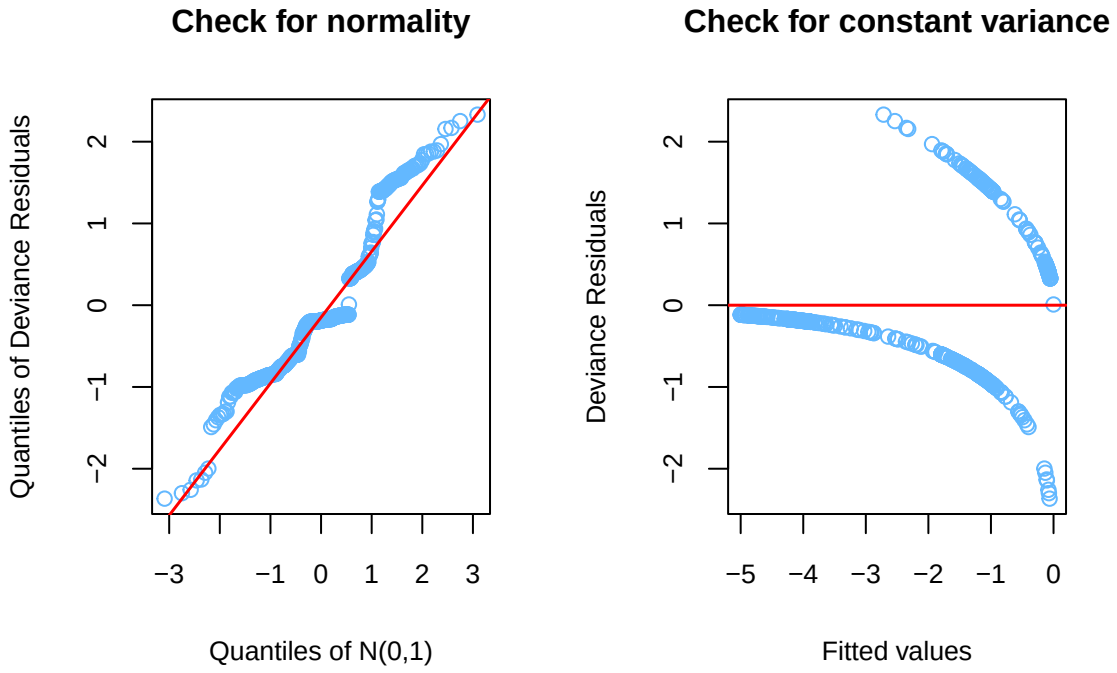


Figure 2: Checking normality and constant variance in the deviance residuals

Based on the model summary, we observe that the p-values for each covariate are very small, indicating that they are statistically significant in the model and serve as good predictors for the probability of a quick sale. Specifically, the coefficient for agent B is 3.3818, which is positive. This implies that, compared to the reference category (agent A), agent B is associated with a significantly higher probability of selling a property quickly. Furthermore, the smooth term `s(price)` has an estimated 8 degrees of freedom, reflecting a non-linear relationship between price and the probability of a quick sale. This term is also highly significant, emphasising that price plays a crucial role in determining quick sale likelihood.

The existence of two distinct agents (agent A and agent B) in the dataset leads to separate fitted trends for each agent, as illustrated in Figure 3. This difference allows for a comparative analysis of agent performance across different price points. In Figure 3, agent B consistently exhibits a higher probability of achieving quick sales compared to agent A across almost all price ranges. At the manager's target price range of 200,000 SD, agent B is shown to have a higher probability of a quick sale. Therefore, to maximise the chances of selling their property within two months, the manager should employ agent B.

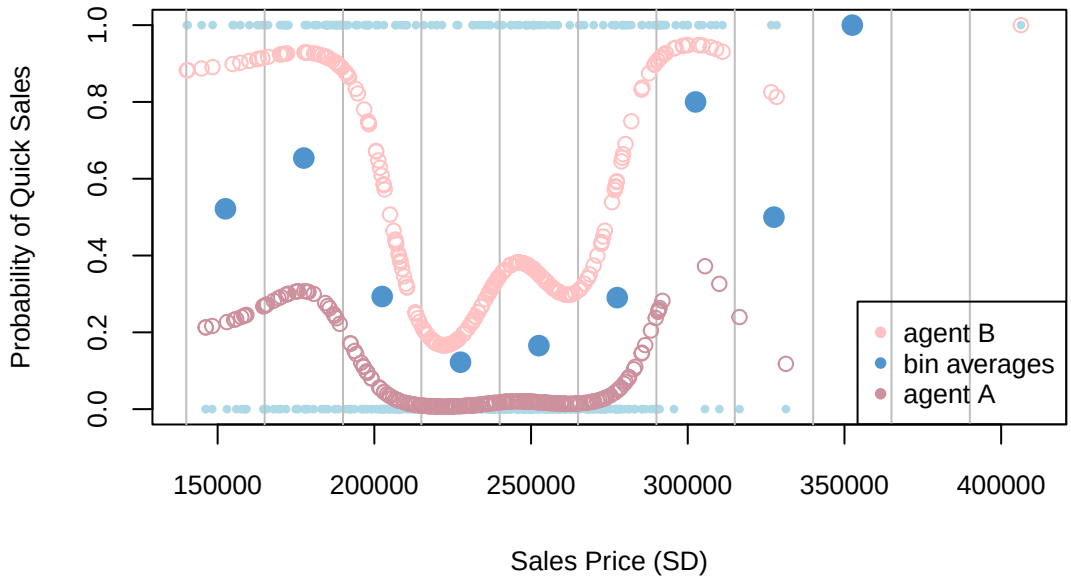


Figure 3: Comparing fitted values and the proportion of quick sales in price bins

In conclusion, both `price` and `agent` are significant predictors in the generalised additive model for `times`. While the adjusted R-squared and deviance explained are moderate, the model still provides valuable insights to assist the manager in selecting the estate agent that is most likely to maximise the probability of achieving a quick sale within 2 months.

**Total word count:*2918*

Statement about use of generative AI tools

We used Deepl to translate some Words and sentences and ChatGPT to refine the grammer mistakes in our assignment.